

Assessing Survival Forests for Prognosis Based on Gene Profiles

Thu M. Hoàng, Université René Descartes
Van L. Parsons^{1,2}, National Center for Health Statistics

Abstract

Combinations of survival regression trees called survival forests (SF) applied to microarray data provide both prediction of individual survival functions and the corresponding ranking of variable importance although without assessment. A basic question is whether a particular SF design and the gene ranking can be attributed to chance alone. For small data sets we propose the use of permutation tests as a way to assess SF, its goodness of prediction and to test the significance of the genes identified as prognostic markers.

1 Introduction

DNA microarrays are a potentially powerful technology for improving prognostic assessment in view of individualized treatment. However this technology delivers very high dimensional noisy data that demand new approaches of statistical analysis to bring patterns to emerge without falling into pitfalls such as overfitting that result in false leads and erroneous conclusions.

A common way to do prognostic analysis using gene profiles is to first cluster samples into classes based on gene expression and perform standard survival analysis to the identified clusters so that genes influencing survival may be uncovered. Alizadeh et al. [1] considered the taxonomy of diffuse large B-cell lymphoma (DLBCL) derived from two-way clustering of genes into functional groups and samples into classes based on gene expression. They demonstrated the existence of two DLBCL patient subgroups with differential survival and showed that this molecular dissection of the disease and the clinical International Prognostic Index identify different features that influence survival. Using hierarchical clustering on 31 cutaneous melanoma cDNA microarrays Bittner et al. [2] identified two groups showing a difference in survival - although not statistically significant (p-value=.13) due to small sample size (n=15) and low event rate (7 deaths). Using hierarchical clustering on oligonucleotide microarrays data, Ferrando et al. [9] identified previously unrecognized molecular subtypes of T-cell acute lymphoblastic leukemia (T-ALL) and showed that activation of the HOX11 oncogene confers a significantly better prognosis as compared to expression of TAL1 and LYL1 oncogenes in terms of patients' survival.

Due to a possible high degree of dependence of the results on the choice of the clustering algorithm, one might not be able to draw valid biological based conclusions for the observed groups. An alternative strategy is then to directly analyze patients' survival using gene expression profiles as features. Standard method for survival analysis is Cox regression assuming proportional hazards. Violations of this assumption in certain applications, conceivably post-genomic analysis of survival based on gene profiles, has led to alternative approaches such as tree-structured survival analysis [19][20][18], neural networks [13],[22] and recently bagging survival trees for increased efficiency [6][14][15] [16].

¹Research results and conclusions expressed in this paper are those of the authors and do not necessarily indicate concurrence by the National Center for Health Statistics.

²The authors contribute equally.

Breiman (2002 Wald lecture [6]) has suggested combinations of survival regression trees called survival forests (SF) for estimating survival functions and importance. Here, a survival tree is grown using a bootstrap sample, and an ensemble of trees is created over many such samples. This ensemble of trees can then be used to estimate survival and infer prognosis for new observations; A new case can be run through the ensemble of trees to obtain an averaged survival function, which in turn gives survival prediction.

In current clinical practice prognoses are based on clinical variables (such as age, health status, size and morphological characteristics of the tumor). While SF can be applied to gene profiles to improve prognostic analysis, the distribution theory of resulting statistics seems intractable. In practical applications, the making of inference based upon a small sample realization becomes problematic. A basic question is whether the SF design and resulting statistics can be attributed to chance alone. We propose the use of permutation tests as a way to determine significance of the SF performance in assessing both the fit/prediction, and the variables importance. We also propose new assessment statistics for SF, and illustrate the techniques by analyzing T-ALL data previously studied by Ferrando [9, 23].

2 Survival Forests

We only provide some highlights of SF, and refer the reader to Breiman [6] for additional details. For a random survival time, $T(\mathbf{x})$, depending on a covariate vector \mathbf{x} , define the hazard function at time t , $h(t, \mathbf{x})$, and the survival function, $S(t, \mathbf{x})$, by

$$\begin{aligned} h(t, \mathbf{x}) &= P(T(\mathbf{x}) \in (t, t + dt) | T(\mathbf{x}) > t) / dt \\ S(t, \mathbf{x}) &= P(T(\mathbf{x}) \geq t) = \exp\left(-\int_0^t h(\tau, \mathbf{x}) d\tau\right) \\ &:= \exp(-H(t, \mathbf{x})) \end{aligned}$$

The data consist of N independent survival times, but subject to random independent right censoring. Such a sample will be represented as $(\mathbf{x}_i, t(\mathbf{x}_i), c_i), i = 1, 2, \dots, N$ where $c_i = 1$ (0) if the survival outcome is observed (censored). A survival regression tree can be grown using maximum likelihood to define splitting rules on time or covariates. The terminal nodes return the levels of a step function that estimates the hazard $h(t, \mathbf{x})$.

While a single tree can be constructed from all the data, estimates based upon single trees may be unstable (e.g., see [13]). For classification and regular regression Breiman [4] introduced the ensemble technique based on bootstrap, called bagging (bootstrap aggregating). Similarly for survival data in presence of censoring Breiman proposes to use a bootstrap sample to create each survival tree and then average over many such trees, hence the name forest, in the hope of reducing variability and improving accuracy. Buhlmann and Yu [7] provide theoretical discussion of bagging.

Here, bagging requires a bootstrap sample of size N from the original N data points to be selected and a survival tree grown. Denoting a single bootstrap sample as b , for each out-of-bag observation (*oob*), i.e., a case not used in the bootstrap sample, the covariate \mathbf{x}_{oob} is put through this single tree to get an estimate of the survival curve, $S_b(t, \mathbf{x}_{oob})$. The bootstrap sample serves as a training set, and training over many bootstrap samples obtains the corresponding trees. The averaged survival function estimated for all the *oob* cases is de facto a test set estimate. If an independent test set is also available, for each new observation the covariate \mathbf{x}_{new} can be put through all the trees to get an estimate of the individual survival curve $\hat{S}(t, \mathbf{x}_{new})$. For a single bootstrap sample and out-of-bag \mathbf{x} the survival tree will produce an estimator of the cumulative hazard, $\hat{H}_b(t, \mathbf{x})$, and the survival function $\hat{S}_b(t, \mathbf{x}) = \exp(-\hat{H}_b(t, \mathbf{x}))$ at selected time points. The aggregated estimators of S and H are $\hat{S}(t, \mathbf{x}) = \sum_{b \in B_{\mathbf{x}}} \hat{S}_b(t, \mathbf{x}) / |B_{\mathbf{x}}|$ and $\hat{H}(t, \mathbf{x}) = \sum_{b \in B_{\mathbf{x}}} \hat{H}_b(t, \mathbf{x}) / |B_{\mathbf{x}}|$, where for each \mathbf{x} , $B_{\mathbf{x}} = \{b : \mathbf{x} \text{ is an } oob \text{ case}\}$ and $|B_{\mathbf{x}}| = \text{number in } B_{\mathbf{x}}$. Using only the *oob* cases reduces overfitting.

Hothorn et al. [16] considered bagging survival trees based on the LeBlanc and Crowley method [20]. For a new observation \mathbf{x}_{new} their method provides the estimated survival curve by computing the Kaplan-Meier curve of all observations identified by the terminal nodes containing \mathbf{x}_{new} .

3 Tuning SF

We have observed that for small data sets some additional tuning of the original algorithm may improve the results [14, 15].

The impact of tuning is highly data specific, but one tuning parameter we added to the original algorithm turned out to frequently lead to results as accurate or better than the original algorithm, but with less computational time. This parameter is the number of covariates selected for determining a node split. Recall in random forests [5] the best split at each node is chosen from m randomly selected predictors from the original set of k predictors independently from node to node. This randomization reduces the correlation among the trees, but maintains the strength of each tree's predictive capability. Random forests requires computing time $\propto \sqrt{m}N \log(N) \times$ number trees, that is a reduction by a factor $\sqrt{\frac{m}{k}}$ when m is used in place of k . We include random subset selection of covariates in the design of a SF.

Another tuning parameter is the probability p_s to split along the covariate or the time. Breiman suggested $p_s \geq 0.50$. If $p_s = 1$, i.e., the covariates are not used for prediction, optimization works on time alone, and the estimators $\hat{S}(t, \mathbf{x})$, while distinct, show little variability from curve-to-curve. If $p_s = 0$, the tree growing optimizes on the covariates \mathbf{x} alone, and the curves $\hat{S}(t, \mathbf{x})$ are based upon a hazard function that is constant in time.

4 Goodness of prediction

The original focus of SF was upon prediction with some corresponding error estimate. For assessment Breiman used only uncensored times. Such approach could lead to biased evaluation when censoring is high. We consider two measures that use both censored and uncensored observations.

Brier integrated score adjusted for censoring. Graf et al. [11] suggest introducing a survival status variable, 0 or 1, with adjustments for censoring to get more accurate evaluation than those obtained by using just the noncensored cases. If $\hat{S}(t|\mathbf{x})$ is the probability of the event $T(\mathbf{x}) \geq t$, the quality of this prediction adjusted for censoring, can be evaluated by using the Brier mean squared error scores

$$B_s(t) = \frac{1}{N} \sum_{i=1}^N \left(\hat{S}(t|\mathbf{x}_i)^2 1_{t_i < t} c_i \hat{G}(t_i)^{-1} + (1 - \hat{S}(t|\mathbf{x}_i))^2 1_{t_i \geq t} \hat{G}(t)^{-1} \right)$$

where $t > 0$, t_i are observed times, and $\hat{G}(t)$ the Kaplan-Meier estimate of the distribution of the time to censor assumed free of \mathbf{x} . If all $c_i = 1$, then $\hat{G} \equiv 1$, and the above is the traditional mean square error. The integrated Brier score for global assessment is

$$B_I = \int_0^{\max(t_i)} B_s(t) dt / \max(t_i).$$

Usually assessment relies on cross-validation or independent test sets. However in small samples, observations are too precious to be set apart for such assessment scheme. In that case the bootstrap sample serves as the training set and the oob observations as a de facto test set. While $\hat{S}(t|\mathbf{x})$ is only based on those trees where \mathbf{x} is oob there remains some dependency among the $\hat{S}(t|\mathbf{x})$ over the different \mathbf{x} since recurrent bootstraps will eventually cover the whole data.

Brier integrated scores may be evaluated for the individual survival curves generated from a SF, $\hat{S}(t, \mathbf{x}_i)$, the average curve, $\bar{S}(t, \cdot) = \sum_{i=1}^N \hat{S}(t, \mathbf{x}_i)/N$ and the Kaplan-Meier estimator. Ratios of the former to either of the latter two scores should be less than one to signify an impact of the covariates. The ratio of the latter two should be about one since the average of the individual curves should be close to a population estimate.

Harrell c-index. The c-index [12] is the proportion of predictions that are concordant out of all pairs of observations for which ordering of the survival times can be determined. Pairs are ignored if the ordering of the true outcomes cannot be determined, i.e., both are censored, or one is censored at a time before the other’s event. Thus, this measure uses both uncensored and censored survival times. It ranges from 0 to 1, and equals 0.5 for the constant predictor. A c-index near 0.5 means that the model is not predictive while a c-index near 1 indicates that the model is highly predictive. From each curve several predictors may be derived, e.g., means, medians, or probability of survival at a time threshold, and a c-index computed for each of them. It is a generalization of the ROC index to censored data.

5 Permutation tests to assess SF

If the c-index > 0.5 or if the Brier score for SF survival curves shows improvement over its counterpart for the Kaplan-Meier curve, one may like to determine its significance in relation to a “null” standard. As of present, SF does not offer provision for hypothesis tests.

5.1 Global Test of Exchangeability

The idea is to compare SF statistics to counterparts obtained from simpler “no-covariate” models, e.g., the Kaplan Meier estimator, or a survival forest using only time as a split variable, i.e., $p_s = 1$. One way to do so is to call upon conditional permutation tests under an hypothesis of exchangeability to evaluate the significance of SF statistics. See Good [10] or Pesarin [21] for a general discussion. While a global null hypothesis $H_0 : h(t|\mathbf{x}) = h(t)$, h free of \mathbf{x} is of interest, we must consider a similar albeit stronger hypothesis of exchangeability which is amenable to permutation tests.

Consider N observations, labelled $i = 1, 2, \dots, N$ with associated covariate vector \mathbf{x}_i and the survival response (t_i, c_i) . The exchangeable hypothesis H_0^e assumes that the distribution of $(\mathbf{x}_i, (t_i, c_i))$ is invariant in any permutation (i_1, i_2, \dots, i_N) of $(1, 2, \dots, N)$. In a practical sense, it means that \mathbf{x} provides no information about survival.

This hypothesis however strong may help quantify the orders of magnitude for test statistics derived from SF for small samples such as the Brier score or the c-index. If $\hat{f}(\mathbf{x}, (t, c))$ is a statistic we can generate its H_0^e distribution by permuting the (t_i, c_i) ’s but keeping \mathbf{x}_i fixed. For each permutation, say π , a complete forest is grown and the statistic $\hat{f}(\mathbf{x}, \pi(t, c))$ computed. Its empirical distribution is obtained over a large number of permutations and can be used to gauge the magnitude of $\hat{f}(\mathbf{x}, (t, c))$. Extreme values of the statistic provide evidence against null hypothesis of exchangeability.

5.2 Permutation Tests for Selecting Important Variables

Important variables for survival To study the relation of a covariate component x_p with survival, Breiman suggests looking at the correlations, ρ , of x_p with an estimator of the cumulative hazard, $\hat{H}(t|\mathbf{x}) = -\log(\hat{S}(t|\mathbf{x}))$ for select time points. Call its estimator $\hat{\rho}(t, x_p) := \rho(\hat{H}(t|\mathbf{x}), x_p)$. Breiman only considered uncensored survival times, but for small samples where these times appear skewed, censored times can be included for evaluating $\hat{H}(t|\mathbf{x})$.

In studies with large numbers of covariates, the many relations among covariates may result in many large absolute estimates of $\hat{\rho}(t, x_p)$. To evaluate their significance, we propose

computing the distribution of $\hat{\rho}(t, x_p)$ under the exchangeability hypothesis, H_0^e , and using 2-sided asymmetric p-values of the observed correlation $\hat{\rho}_{obs}$, i.e., $2 \min(P_{H_0^e}(\hat{\rho}(t, x_p) \leq \hat{\rho}_{obs}), 1 - P_{H_0^e}(\hat{\rho}(t, x_p) \leq \hat{\rho}_{obs}))$ to allow for eventual skewness.

Small p-values mean evidence against H_0^e , but the ordering of p-values does not necessarily rank the covariates. Care must be exercised in assessment.

Multiple hypothesis testing In the past decade methods that improve upon the highly conservative Bonferroni procedure and implicitly incorporate the dependence structures of the test statistics have been developed. Dudoit et al [8] give a comprehensive discussion. Table 2 in that reference provides assumptions needed to apply some of the newer methods. Given the success of using the step-down min P and max T methods with permutation t-type tests, we had hoped they could help distinguish important covariates based upon $\hat{\rho}(t, x_p)$. Unfortunately, for the T-ALL example and most likely for many microarray data sets with concurrent survival information, the null structure did not conform with the needed assumption of subset pivotality. It appears that only the more conservative procedures may be adapted for multiple testing.

5.3 Limitations on permutation tests

Hsing et al. [17] addressed issues as to whether permutation p-values of error estimates are informative in comparing different classifiers. For models much simpler than SF, they found that p-values are slowly increasing functions of the error estimates, and close to 0 and somewhat flat for small or moderate error rates. Thus, relating the magnitudes of error estimates to p-value may be problematic in practice. Hsing concluded that it is possible for p-values to be less informative than the error estimates. Our results for SF applied to T-ALL microarray data are consistent with Hsing's observation.

Here the tuning parameters m and p_s were selected using the Brier score and c-index. However of moderate utility for designing a SF the permutation tests help in providing quantitative assessment of covariate importance after the tuning parameters have been selected. While the permutation statistics seem reasonable, their power has not been studied.

6 T-ALL Example

To demonstrate the techniques of the sections 4 and 5 to help choosing m , and p_s , and to select significant genes for survival prediction we applied to T-ALL data from Ferrando et al. [9]. The data consisting of 39 T-ALL samples were analyzed with both DNA microarray (Affymetrix HU6800 with 7129 probe sets) for the global patterns of gene expression and RT-PCR (reverse transcriptase polymerase chain reaction) for expression of single genes. RT-PCR detected 29 samples with aberrant expression of the oncogenes HOX11, LYL1 or TAL1, and 10 without detectable expression of these oncogenes. By permutation tests Ferrando et al. obtained 72 genes whose expression patterns best distinguished the 4 phenotypes. Then using these genes they clustered the samples and identified 3 main tumor classes and 2 tumor subclasses. When estimating Kaplan-Meier survival curves for the cases within each cluster Ferrando et al. obtained similar results for both the RT-PCR and micro-array methods. A follow-up of this study [23] considered the role of cyclin D3 in leukemogenesis and reanalyzed the differential expression of the genes clustered with cyclin D3 in k-nearest neighbor clustering. In our analysis we concatenated the two lists of genes derived from the two analysis to get 79 genes.

Ferrando et al. did *not* use the survival times to select the genes. A secondary check on 79 univariate Cox regressions for the 39 samples found 41 of the selected genes had significance greater than 0.25, thus confirmed that the preliminary selection was independent of the survival. The censoring rate is high, thanks to good prognosis of T-ALL in children. Only 12 survival times were uncensored and the 13 largest times were censored. Furthermore, 4 survival times were 0 which we recoded by adding 1 month.

For our analysis the intention was to start with the 39 subjects and 79 genes and let SF separate the individuals by survival curves and separate the genes by the correlations $\hat{\rho}(t, x_p)$. Since the purpose of this research is to gain a better understanding of the permutation tests for assessing SF we do not attempt to make any final substantive biological inference.

6.1 Designing SF

SF's were grown with 100,000 bootstraps, $m = 1, 2, 4, 8, 16, 24, 32, 64, 79$ for random selection among the 79 genes and $p_s \in [0.10, 0.90]$. For each m , the same bootstrap sample was used to reduce sampling variability when making comparisons among the tuning parameters. Several modifications to the original algorithm were made, but we will restrict our discussion to the two parameters, m and p_s , that seem to have the broadest applications. A comparison of the Brier scores and c-indexes for different choices of (m, p_s) showed no distinct optimal values, but intervals of optimal performance. Selecting $m \in [8, 24]$ and $p_s \in [0.30, 0.50]$ resulted in somewhat flat Brier scores and c-indexes. Designing SF was robust with respect to the choice of (m, p_s) . In practice, $m = 16$ and $p_s = .5$ seem adequate.

Estimating the null distribution of a statistic requires a large number of SF's to be simulated, each based upon a permutation of the survival and censored times. To grow 1000 distinct forests for a fixed m we reduced the number of bootstraps per forest to 20,000.

The p-values were rather flat in the parameter ranges considered above, and $p \in [0.015, 0.02]$ for the Brier scores, $p \approx .06$ for the c-indexes at time points 36, 60 and 142 months, and $p \approx .15$ for the c-index at 12 months. Overall, they suggest modest significance of the Brier and c-indexes values when compared to a "null" standard.

6.2 Selecting significant genes for T-ALL survival

Once the evidence against H_0^c is acquired, the next step is to determine which genes have influence on survival. Significance of $\hat{\rho}(t, x_p)$ at time points 12, 36, 60 and 142 are given in Table 1. A positive sign for ρ means a positive association with hazard. As mentioned earlier, no adjustment for multiple testing has been used, and care must be taken in any comparison of the $\hat{\rho}(t, x_p)$'s by magnitude or p-value ordering. While $\hat{\rho}(t, x_p)$ is a "standardized measure" the marginal null distributions appear quite distinct and joint inference becomes difficult. It is of interest to see that some genes are related to survival across the whole time range, notably genes 52 and 32, while others are important for long term survival such as genes 07 and 48. For comparison, the asymptotic p-values from univariate Cox regressions have been included. Note that the values of $\hat{\rho}(t, x_p)$'s agree in sign with the Cox results. Some genes appear significant for the SF model but not Cox model, e.g., gene 49 which turned out to be expressed in the HOX cases known to have better prognosis.

7 Discussion

We have demonstrated techniques for assessing fit and making inference when using SF on small sample problems. The Brier integrated score and c-index are useful in tuning and assessing SF for goodness of prediction. SF performance appears robust with respect to tuning parameters : changes in tuning result in little change in the fit and prediction.

In fitting models or establishing important variables, the significance of an observed statistic by chance alone is often of interest. We suggest using p-values generated from permutation tests for SF. For model selection our results are consistent with previous observation by Hsing et al. that p-values may be less informative than error measures. However, p-values can provide the coarse information whether a model is unlikely by pure chance. The p-values for $\hat{\rho}(t, x_p)$ appear more informative. SF curves produced with many highly correlated predictors will have many large $|\hat{\rho}(t, x_p)|$, and the p-values may help to sepa-

Table 1: The most important genes for the hazard or the survival at 4 selected time points according to p-value of $\hat{\rho}(t, x_p)$ with significance level at 0.05. A positive sign indicates positive correlation with hazard, and a negative sign indicates positive correlation with survival.

gene	$t = 12$	p-val	$t = 36$	p-val	$t = 60$	p-val	$t = 142$	p-val	Cox p-val
	ρ		ρ		ρ		ρ		
52	0.67	0.003	0.69	0.001	0.69	0.001	0.69	0.001	0.02
32	0.75	0.005	0.80	0.001	0.80	0.001	0.81	0.001	0.06
16	-0.53	0.023	-0.54	0.021	-0.55	0.019	-0.56	0.017	0.07
18	0.65	0.029	0.66	0.031	0.68	0.023	0.69	0.011	0.03
51	0.39	0.035	0.43	0.031	0.42	0.039	0.42	0.045	0.06
66	-0.46	0.037	-0.50	0.023	-0.51	0.025	-0.51	0.029	0.00
49	-0.38	0.041	-0.42	0.029	-0.43	0.027	-0.44	0.029	0.75
61	0.52	0.043	0.58	0.019	0.56	0.033	0.54	0.043	0.26
44	0.49	0.045	0.52	0.031	0.51	0.031	0.51	0.033	0.31
19	0.49	0.053	0.55	0.023	0.55	0.027	0.54	0.037	0.06
56	0.66	0.069	0.64	0.095	0.68	0.063	0.69	0.043	0.02
07	0.62	0.081	0.67	0.031	0.67	0.031	0.67	0.033	0.03
48	0.62	0.097	0.71	0.019	0.70	0.037	0.68	0.047	0.20
27	0.51	0.117	0.62	0.033	0.61	0.043	0.60	0.053	0.00

rate them. The null distributions, however are not necessarily comparable, and currently available less conservative multiple testing procedures do not seem directly applicable.

As discussed earlier, SF computing time is $\propto \sqrt{m}N \log(N) \times$ number trees. Practical computing constraints may limit the feasibility of using permutation tests whenever any of the three components grows large in size.

References

- [1] Alizadeh, A.A., Elsen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Ran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503-511.
- [2] Bittner M., Meltzer P., Chen Y., Jiang Y., Seftor E., Hendrix M., Radmacher M., Simon R., Yakhini Z., Ben-Dor A., Sampas N., Dougherty E., Wang E., Marincola F., Gooden C., Lueders J., Glatfelter A., Pollock P., Carpten J., Gillanders E., Leja D., Dietrich K., Beaudry C., Berens M., Alberts D. and Sondak V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406: 536-540.
- [3] Breiman L., Friedman, J.H., Olshen R.A. and Stone C.J. (1984) *Classification and regression trees*. Wadsworth, Belmont, CA.
- [4] Breiman, L. (1996) Bagging predictors, *Machine Learning* 24, 123-140.
- [5] Breiman, L. (2001) Random Forests, *Machine Learning* 45, 5-32.
- [6] Breiman, L. (2002) Wald III Lecture: Software for the masses, *Lecture notes available at* <http://stat-www.berkeley.edu/users/breiman/wald2002-3.pdf>
- [7] Buhlmann P., Yu B. (2002) Analyzing bagging, *The Annals of Statistics* 30, 927-961
- [8] S. Dudoit, J. Popper Shaffer and J. C. Boldrick (2003) Multiple Hypothesis Testing in Microarray Experiments, *Statistical Science*, Vol. 18, No. 1, 71-103

- [9] Ferrando A.A., Neuberger D.S., Staunton J., Loh M.L., Huard C., Raimondi S.C., Behm F.G., Pui C.H. Downing J.R., Gilliland D.G., Lander E.S., Golub T.R. and Look A.T. (2002) Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia *Cancer Cell*, 1, 75-87
- [10] Good, P. (2000), *Permutation tests: a practical guide to resampling methods for testing hypotheses* Berlin, Springer-Verlag.
- [11] Graf Z., Smchour C., Sauerbrei W., and Schumacher M. (1999) Assessment and comparison of prognosis classification for survival data *Statistics in Medicine*, 18, 2529-2543
- [12] Harrell FE., Lee KL. and Mark DB.(1996) Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine* 15 361-387.
- [13] Hoàng T. Trinh QA. and Asselain B. (2002) Construction and validation of a prognostic model for metastatic breast cancer using Bayesian neural network and regression tree, *Intelligent Data Analysis in Medicine and Pharmacology, Workshop Notes*, 37-43 <http://www.cs.uu.nl/lucas/idamap2002/idamap2002-proc.pdf>.
- [14] Hoàng T. and Parsons V. (2004) Breast Cancer Studies Breast Cancer Prognosis using Survival Forests, in *Reliability, Survival Analysis, and Quality of Life Series* Statistics for Industry and Technology Nikulin, M.S.; Balakrishnan, N.; Mesbah, M.; Limnios, N. (Eds.) to appear, Birkhauser
- [15] Hoàng T. and Parsons V. (2004) Bagging Survival Trees for Prognosis based on Gene Profiles, To appear *Compstat'2004* Physica Verlag
- [16] Hothorn T., B. Lausen , A. Benner and M. Radespiel-Troger (2004) Bagging survival trees, *Statist. Med.* ; 23:77- 91
- [17] Hsing T. Attoor S. Dougherty E. (2003) Relation Between Permutation-Test P Values and Classifier Error Estimates *Machine Learning*, 52, 1130.
- [18] Intrator O. and Kooperberg C. (1995) Trees and splines in survival analysis, *Statistical Methods in Medical Research* 4, 237-261
- [19] Keles S. and Segal MR. (2002) Residual-based tree-structured survival analysis, *Statistics in Medicine* 21, 313-326
- [20] LeBlanc M. and Crowley R. (1992) Survival trees by goodness of fit, *Journal of the American Statistical Association* 88, 457-467
- [21] Pesarin, F (2001), *Multivariate permutation tests: with applications in biostatistics*, John Wiley & Sons New York, Chichester
- [22] Ripley R. M., Harris A. L., Tarassenko L.(2004) Non-linear survival analysis using neural networks *Statistics in Medicine* 23(5) 825-842
- [23] Sicinska E., Aifantis I., Le Cam L., W. Swat, C. Borowski Q. Yu A. A. Ferrando S. D. Levin, Y. Geng H. von Boehmer and P. Sicinski (2003) Requirement for cyclin D3 in lymphocyte development and T cell leukemias *Cancer Cell* (4) 451-461